Biometrical Letters
Vol. 48 (2011), No. 2, 113-121

Application of statistical tests in gene selection problems

Joanna Zyprych-Walczak, Alicja Szabelska, Idzi Siatkowski

Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland, e-mail: zjoanna@up.poznan.pl, aszab@up.poznan.pl, idzi@up.poznan.pl

SUMMARY

Selection of genes is an important issue in discriminant analysis. In this paper we present the use of some statistical tests. Several existing statistical methods such as the F-test, Kruskal-Wallis test for testing the equality of means and Bartlett test, Fligner-Killeen test, and Levene test for testing homogeneity of variance, are presented and compared. These techniques make it possible to find sets of significant genes with different efficiency in relation to discriminant analysis. We present the results obtained based on misclassifications of samples derived with usage of lists of significant genes obtained for the considered tests.

Key words: gene, R software, selection, statistical test.

1. Introduction

The technology of microarrays allows the investigation of thousands of genes at the same time. It enables one to determine information about the expression profile of genes. Statistical analysis is widely used in searching for over- and under-expressed genes. Apparently, there exist many statistical tests for verifying hypotheses. The classic example of such procedures is a group of tests verifying the equality of means of expression levels. The researcher can often be unsure as to the choice of the most appropriate test in a given investigation. This paper provides assistance in solving this problem. Firstly, within the group of tests verifying the equality of means, an analysis of the efficiency of these tests is performed with respect to classification of differentially expressed genes. Secondly, an analogous analysis is undertaken for tests concerning the equality of variances. Thirdly, based on the previously selected genes as a training

set, the prediction of the chosen sample with remaining genes is tested, applying several methods of machine learning techniques. As the results of the analysis we present the values of misclassified samples. The aim of this paper is to compare several statistical tests and review the usefulness of these tests in the selection of genes from microarray experiments.

We would like to note that all the computations were performed with the use of the R platform, version 2.10.0 (R Development Core Team, 2009).

2. Data

In the analysis we consider data with more than two classes. Data are presented as a matrix where each row contains a gene and each column a sample of mRNA. Hence the values of these matrices are the expression levels of genes for given samples. In the case of the first dataset – 'leukemia 72' (Golub et al., 1999) – three classes of features were specified. These data contain the expression levels of 7129 genes that were jointly examined in 72 samples. The next dataset – 'ovarian' (Dudoit et al., 2002) – has 39 samples with expression levels of 7129 genes in each of these samples. The last dataset analyzed in this paper – 'lung cancer' (Hartung et al., 2002) – presents the expression levels of 918 genes in 73 samples. The use of datasets known from the literature was deliberate, as this makes it possible to compare our results with those obtained in other published papers. In addition, we extend our analysis to investigate more tests compared with Welsh et al. (2001) or Dechang Chen et al. (2005).

3. Methodology and purpose

Selection of genes focuses on identifying genes which are differentially expressed in analyzed groups of samples. The statistical tests that are used for determining such genes can be divided into two groups. The first group includes tests that analyze the relevant differences between the mean levels of expression between several groups of genes, e.g. the F-ANOVA test and Kruskal-Wallis test, whereas the second group consists of tests that investigate homogeneity of variance for several groups of genes,

e.g. Bartlett test, Fligner-Killeen test and Levene test. When each gene is tested separately, one p-value per gene is obtained. The actual p-value was verified with use of the FDR correction based on the procedure introduced by Benjamini and Hochberg (1995). In both groups of tests, we chose the genes for which the adjusted p-values are below 0.05 according to the appropriate test. These genes were used to determine the ranks of the differentially expressed genes in the subsequent analysis.

The selection of genes was performed with use of the five tests mentioned above, based on the three considered datasets. For every test, the values of expression levels of each gene in several groups were considered. It was evaluated whether the data differ significantly between groups. Next, genes were ranked with respect to the adjusted p-values. From every sample there were selected, respectively, 50, 100 and 200 of the most differentially expressed genes. The chosen sets of genes were subjected to three prediction methods: naïve Bayesian method (NB), k-nearest neighbor method (KNN) and support vector machine method (SVM) (Krzyśko at al., 2009). Cross validation (leave-one-out cross validation) was performed for the classifier obtained with the use of one of these methods. The analyzed set of data was divided into training and tested sets. The classifier is constructed with the use of the first set. In the leave-one-out cross validation the training set contains of n-1 data points, where n is the number of samples. Cross validation was performed based on the 50, 100 and 200 most differentially expressed genes obtained from the considered tests. For clarity, in Tables 1–3 and Figures 1–2 we present the results based on 100 (for Tables) and 50 (for Figures) genes only. Next, the classifier is tested based on the remaining one data point. This procedure is repeated for every data point in the set. At each step of the calculations we determine the error which identifies whether the remaining data point was correctly classified. As a result we obtained the number of misclassified samples based on the chosen classifier. Next the errors of prediction were compared for every test mentioned above and for the three prediction methods.

Table 1. Percentage intersection of the tests for the dataset 'leukemia72' based on the 100 most significant genes analyzed.

TESTS	F-ANOVA vs. Kruskal-Wallis		
Intersection		54%	
TESTS	Bartlett vs Fligner-Killeen	Bartlett vs Levene	Fligner-Killeen vs Levene
Intersection	45%	28%	60%

Tables 1, 2 and 3 present the intersection of the analyzed dataset determined based on a comparison of the considered tests. Percentage values were verified for the joint part of the 100 genes that were identified as the most informative ones. Results for the tests analyzing equality of means and homogeneity are shown separately. The bold values in the tables signify the tests for which the joint parts were the largest.

Table 2. Percentage intersection of the tests for the dataset 'ovarian' based on the 100 most significant genes analyzed

TESTS	F-ANOVA vs Kruskal-Wallis		
Intersection		65%	
TESTS	Bartlett vs Fligner-Killeen	Bartlett vs Levene	Fligner-Killeen vs Levene
Intersection	22%	11%	63%

Table 3. Percentage intersection of the tests for the dataset 'lung cancer' based on the 100 most significant genes analyzed

TESTS	F-ANOVA vs Kruskal-Wallis		
Intersection		69%	
TESTS	Bartlett vs Fligner-Killeen	Bartlett vs Levene	Fligner-Killeen vs Levene
Intersection	43%	45%	83%

A summarization of the joint portion of genes is presented in the following Venn diagrams. The first three diagrams (Fig. 1) concern the tests analyzing equality of means, while the other three (Fig. 2) concern tests of homogeneity based on the 50 most significant genes analyzed.

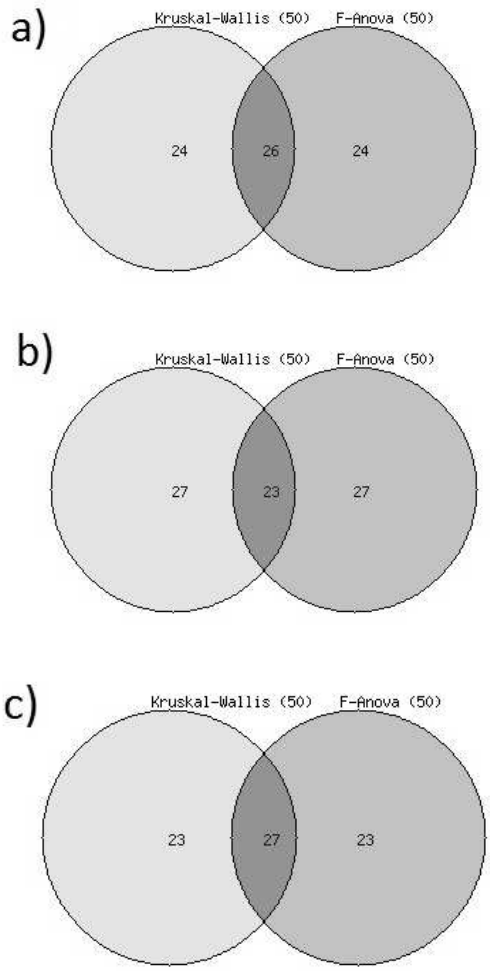

Figure 1. Venn diagram for tests verifying the equality of means in the case of a) 'lung cancer' data, b) 'leukemia72' data, c) 'ovarian' data, based on the 50 most significant genes analyzed.

Tables 4, 5 and 6 show the results of computations (number of misclassified samples) for the datasets 'leukemia72', 'ovarian' and 'lung cancer' respectively, for the 5 considered tests and 3 methods of prediction.

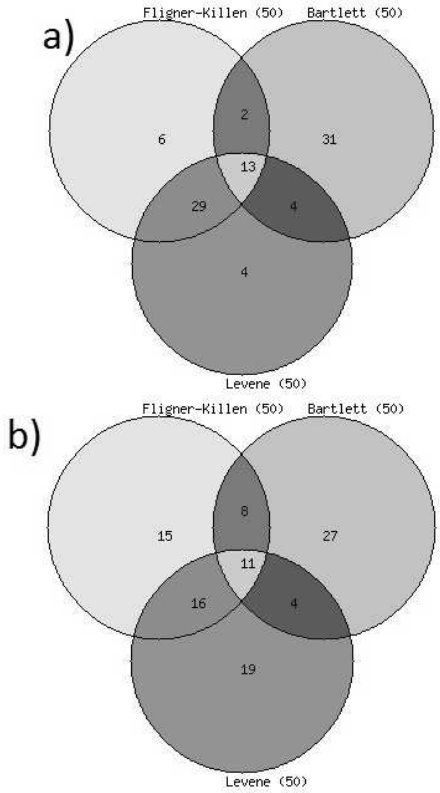

Figure 2. Venn diagram for tests verifying the homogeneity of a) 'lung cancer' data, b) 'leukemia72' data, c) 'ovarian' data, based on the 50 most significant genes analyzed

4. Conclusions

Tables 1–3 and all the Venn diagrams were compiled for comparison of the tests. In the group of tests analyzing equality of means it was observed that the intersection of genes for the Kruskal-Wallis and F-ANOVA tests have respectively 54%, 65% and 69% joint informative genes considering the 100 most informative genes for each dataset. With the 50 most informative genes, 52%, 46% and 54% joint informative genes respectively were identified for each dataset. Tests of homogeneity reveal that the Levene test has at least 60% joint informative genes out of 100 with the Fligner-Killeen test, and at least 54% joint informative genes out of 50 for each dataset. Tests analyzing equality

Table 4. Number of misclassified samples for the dataset 'leukemia72'

Statistical test	Prediction method	50 genes	100 genes	200 genes
	1. NB	3	3	1
Kruskal-Wallis	2. KNN	4	7	3
	3. SVM	2	4	3
AVERAGE ERROR		3	4.7	2.3
	1. NB	1	1	1
F-ANOVA	2. KNN	3	5	4
	3. SVM	3	3	3
AVERAGE ERROR		2.3	3	2.7
	1. NB	2	3	1
Fligner-Killeen	2. KNN	6	5	3
	3. SVM	7	4	4
AVERAGE ERROR		5	4	2.7
	1. NB	3	3	0
Bartlett	2. KNN	6	5	3
	3. SVM	9	10	13
AVERAGE ERROR		6	6	5.3
	1. NB	1	0	0
Levene	2. KNN	4	3	3
	3. SVM	2	1	1
AVERAGE ERROR		2.3	1.3	1.3

of means show the lowest error of misclassification. The error of prediction for these two tests was very often minimal compared with the other tests. Although the most popular method of gene selection is the equality of means, our results show that in some cases tests of homogeneity outperform the former tests.

In addition, the Bartlett test resulted in the lowest number of joint genes with the other tests. This test gives the highest prediction error in every case. In particular, the results of this test combined with the SVM method show the maximal error for every dataset.

Table 5. Number of misclassified samples for the dataset 'ovarian'

Statistical test	Prediction method	50 genes	100 genes	200 genes
	1. NB	0	0	0
Kruskal-Wallis	2. KNN	7	7	9
	3. SVM	2	0	2
AVERAGE ERROR		3	2.3	3.7
	1. NB	1	1	0
F-ANOVA	2. KNN	10	12	11
	3. SVM	0	0	0
AVERAGE ERROR		3.7	4.3	3.7
	1. NB	0	0	1
Fligner-Killeen	2. KNN	12	12	6
	3. SVM	0	0	3
AVERAGE ERROR		4	4	3.3
	1. NB	7	5	5
Bartlett	2. KNN	16	12	13
	3. SVM	18	15	15
AVERAGE ERROR		13.7	10.7	11
	1. NB	3	2	2
Levene	2. KNN	11	9	10
	3. SVM	5	4	2
AVERAGE ERROR		6.3	5	4.6

Acknowledgement

The authors would like to thank the reviewer for their constructive suggestions and comments which resulted in a much improved article.

REFERENCES

Benjamini Y., Hochberg Y. (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach Multiple Testing. J. Royal Stat. Society B 57 (1): 289–300.

Dechang Chen, Zhenqiu Liu, Xiaobin Ma, Dong Hua (2005): Selecting Genes by Test Statistics. Journal of Biomedicine and Biotechnology 2: 132–138.

Dudoit S., Fridlyand J., Speed T.P. (2002): Comparison of discrimination methods for the classification of tumors Rusing gene expression data. J. Amer. Statist. Assoc. 97 (457): 77–87.

Table 6. Number of misclassified samples for the dataset 'lung cancer'

Statistical test	Prediction method	50 genes	100 genes	200 genes
	1. NB	17	13	14
Kruskal-Wallis	2. KNN	13	13	11
	3. SVM	12	10	12
AVERAGE ERROR		10.7	12	12.3
	1. NB	16	14	15
F-ANOVA	2. KNN	15	11	9
	3. SVM	10	8	9
AVERAGE ERROR		13.7	11	7.7
	1. NB	16	12	14
Fligner-Killeen	2. KNN	11	14	10
	3. SVM	12	11	12
AVERAGE ERROR		13	12.3	12
	1. NB	19	15	15
Bartlett	2. KNN	19	13	13
	3. SVM	19	19	16
AVERAGE ERROR		19	15.7	14.7
	1. NB	17	14	13
Levene	2. KNN	17	14	11
	3. SVM	16	16	11
AVERAGE ERROR		16.6	14.6	11.6

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439): 531–537.

Hartung J., Argac D., Makambi K.H. (2002): Small sample properties of tests on homogeneity in oneway ANOVA and meta-analysis. Statist. Papers 43:197–235.

Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008): Systemy uczące się. Rozpoznawanie wzorców. Analiza skupień i redukcja wymiarowości. WNT, Warszawa.

R Development Core Team (2009): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

Welsh J.B., Zarrinkar P.P., Sapinoso L.M., Kern S.G., Behling C.A., Monk B.J., Lockhart D.J., Burger R.A., Hampton G.M. (2001): Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA 98(3): 1176–1181.